DOCUMENT RESUME

ED 468 716 TM 034 348

AUTHOR	Wang, Xiang Bo
TITLE	On Giving Test Takers a Choice among Constructive Response Items. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-R-96-03
PUB DATE	1999-03-00
NOTE	32p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Adaptive Testing; Advanced Placement; Chemistry; *College Entrance Examinations; Computer Assisted Testing; *Constructed Response; Essay Tests; *High School Students; High Schools; *Law Schools; *Secondary School Teachers; Test Format
IDENTIFIERS	*Choice Behavior; *Law School Admission Test

ABSTRACT

The Law School Admission Council (LSAC) is currently investigating the feasibility and advisability of administering a computerized Law School Admission Test (LSAT). In this context, using data from the College Board's 1989 National Advanced Placement (AP) Chemistry Examination for 18,462 test takers and a survey of all AP Chemistry teachers in Hawaii, this study investigated the relationship among the essay choices made by national test takers on five essay items, the test takers' ability levels, AP chemistry curriculum, choosing methods, and performance. Major findings are: (1) the five essays under investigation were chosen in dramatically different ways; (2) the more frequently chosen essays belonged to the core chemistry content while the least frequently chosen item addressed a highly specialized chemistry topic; (3) there was a negative correlation between the popularity of essays and their mean scores; (4) the order in which the essays were presented seemed to have a significant effect on choice patterns of all test takers; (5) across the entire ability range, test takers who chose items selectively performed significantly higher than those who chose sequentially, possibly due to fatigue or lethargy; and (6) except for the extremely low-ability test takers, all test takers of all different ability levels seemed to choose in a similar way. An appendix contains the chemistry teacher survey. (Contains 8 tables, 16 figures, and 16 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 468 716

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. VASELECK
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

On Giving Test Takers a Choice Among Constructive Response Items

Xiang Bo Wang
Law School Admission Council

Law School Admission Council
Computerized Testing Report 96-03
March 1999

TM034348

A Publication of the Law School Admission Council

BEST COPY AVAILABLE

LSAC RESEARCH REPORT SERIES

On Giving Test Takers a Choice Among Constructive Response Items

Xiang Bo Wang
Law School Admission Council

Law School Admission Council
Computerized Testing Report 96-03
March 1999

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

LSAT®; *The Official LSAT PrepTest®*; *LSAT: The Official TriplePrep®*; and the Law Services logo are registered marks of the Law School Admission Council, Inc. Law School forum is a service mark of the Law School Admission Council, Inc. *LSAT: The Official TriplePrep Plus*; *The Whole Law School Package*; *The Official Guide to U.S. Law Schools*, and ***LSACD*** are trademarks of the Law School Admission Council, Inc.

Copyright© 1999 by Law School Admission Council, Inc.

All rights reserved. This book may not be reproduced or transmitted, in whole or in part, by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary 1

Abstract 2

Introduction 2
Research Instrument and Data Collection 3

Analyses and Results 4
Phase I: Overall Essay Choice Tendencies and Performance 5
Phase II: Curricular Explanations of Differential Essay Choices 11
Phase III: Relationship Among Test Taker Essay Choices, Ability Profiles, and Performance 13
Phase IV: Exploring the Effect of Sequential vs. Selective Choosing on Performance 20

Conclusions and Discussions 24

References 25

Appendix 26

Executive Summary

Increased emphasis on assessing the "generative" or "constructive" process of learning has more and more testing agencies considering or incorporating constructive response (CR) items into their tests such as writing essays, answering holistic questions, and/or hands-on projects. Administering CR items requires achieving a balance among three competing factors—the limited time usually allowed for taking a test, the scope and depth of content coverage required by CR item specifications, and the extensiveness on the part of the test taker in answering CR items. The common compromise is to allow test takers to choose a subset of CR items that covers a wide content range. Examples of such testing practice are the College Board's Advanced Placement Examinations of chemistry, physics, history, and many other such tests.

However, a review of the few available studies that investigated the consequences of allowing choices on a test has revealed a few disturbing facts. First, due to test security reasons, very few CR items on nationally administered tests are as vigorously pretested as their multiple-choice counterparts with respect to their difficulties and other psychometric properties. Second, because of the lack of preoperational statistical evaluation, many CR items function very differently from one another in their degree of difficulty and discrimination. Third, in many cases, test takers of different gender and ethnic groups seem to choose differently, which results in score biases to their disadvantage. The objective of this paper is threefold—(1) to investigate how examinees choose CR items; (2) to assess the relationship between item choice and test performance; and (3) to understand why examinees choose CR items in the ways they did.

Based on the national data of the 1989 Advanced Placement (AP) Chemistry Examination of the College Board, six major findings have been obtained: (1) Like previous studies, the essays that test takers could choose from differed substantially in both overall and conditional difficulty. They were supposed to be equivalent. (2) The essays were dramatically differentially chosen, while the issue of equity requires similar choice frequencies. (3) The order in which the essays were presented in the test had a significant effect on examinee choice patterns. (4) The examinees who chose items selectively performed significantly higher than those who chose sequentially, regardless of actual ability. (5) Negative correlations were found between examinee performance and choice frequencies—examinees tended to score lower on the more popularly chosen items on which they had expected to perform better. (6) Unequal item difficulty and choices resulted in differential impacts on the performance of examinees of the same ability, with middle-ability examinees affected the most. This threatens the very integrity of any test that allows for choices.

Based on the survey of all the AP chemistry teachers in Hawaii, some explanations of the above findings were obtained: (1) The teachers agreed significantly with the national data on the popularity of the essays. (2) The reason for the differential essay choices is that the more frequently chosen essays belonged to the core chemistry content, while the less frequently chosen items represented special chemistry topics. (3) The differential essay choices were caused by differential textbook coverage and examinee familiarity with the content of the essays. (4) The phenomenon that examinees performed better on the less frequently chosen essays was due to the fact these essays required less knowledge, because they were less emphasized by the AP chemistry curriculum. (5) The reason that test takers who made deliberate choices performed better than those who chose sequentially is—the process of conscious choice maximized their academic strength, while minimizing weakness. (6) A substantial amount of random choosing could be largely attributed to fatigue or boredom, because choice essay sections were placed toward the end of the test.

Law School Admission Council (LSAC) is currently investigating the feasibility and advisability of administering a computerized Law School Admission Test (LSAT). Providing test takers with the option of choosing from among a set of items is one possible new format that could be considered within a computer delivered assessment system. One example may be letting test takers choose from among a list of reading comprehension or writing sample topics. Although the implications of providing test takers with choices are still not fully understood, findings from this study on how test takers make such choices, and how such choices impact their performance can certainly enrich our knowledge and provide us with useful guidelines for possible implementation of similar testing practices.

Abstract

During the past five years, there has been increased emphasis on assessing the "generative" or "constructive" process of learning. More and more testing agencies have started to incorporate constructive response items into their tests, a subset of which might be chosen by test takers.

Law School Admission Council (LSAC) is currently investigating the feasibility and advisability of administering a computerized Law School Admission Test (LSAT). Providing test takers with the option of choosing from among a set of items is certainly possible within a computer delivered assessment system. For example, in the future, LSAT takers might be given the opportunity to choose from among a list of reading comprehension and writing sample topics. However, issues such as how test takers make such choices or how such choices impact their performance have only recently been discussed. The need to understand the implications of providing test takers with choices is the impetus for this study.

Based on the data from the College Board's 1989 National Advanced Placement (AP) Chemistry Examination (The College Board, 1990) and a survey of AP chemistry teachers in Hawaii, this study investigated the relationship among the essay choices made by national test takers on five essay items, the test takers' ability levels, AP chemistry curriculum, choosing methods, and performance. Major findings are: (1) the five essays under investigation were chosen in dramatically different ways; (2) the more frequently chosen essays belonged to the core chemistry content, while the least frequently chosen item addressed a highly specialized chemistry topic; (3) there was a negative correlation between the popularity of essays and their mean scores; (4) the order in which the essays were presented seemed to have a significant effect on choice patterns of all test takers; (5) across the entire ability range, test takers who chose items selectively performed significantly higher than those who chose sequentially, possibly due to fatigue or lethargy; and (6) except for the extremely low-ability test takers, all test takers of all different ability levels seemed to choose in a similar way.

Introduction

During the past five years, there has been increased emphasis on assessing what is called the "generative" or "constructive" process of learning (Wittrock & Baker, 1992). More and more testing agencies have started incorporating constructive response (CR) items into their tests such as writing essays, answering holistic questions, and/or hands-on projects. Examples are the College Board's Advanced Placement Examinations of chemistry, physics, history, and many other such tests. Given competing factors such as the limited time usually allowed for taking a test, the scope and depth of content coverage required by test specifications, and the extensiveness on the part of the test taker in answering CR items, a key component of tests containing CR items is to allow test takers to choose a subset of CR items that covers a wide content range.

LSAC is currently investigating the feasibility and advisability of administering a computerized LSAT. Providing test takers with the option of choosing from among a set of items is one possible new format that could be considered within a computer delivered assessment system. One example may be letting test takers choose from among a list of reading comprehension or writing sample topics. To date, the implications of providing test takers with these choices are still not fully understood as to how test takers make such choices, and how such choices would affect their performance. Therefore, it is necessary to thoroughly investigate the relationship among test taker choices, ability levels, curricular factors, and performance.

A review of the few available studies that have investigated the consequences of allowing CR choices on a test has revealed a few disturbing facts. First, due to test security reasons, very few CR items on nationally administered tests are as vigorously pretested as their multiple-choice counterparts with respect to their difficulty and other psychometric properties. Because of the lack of preoperational statistical evaluation, many CR items function very differently from one another in their difficulties and discrimination (Fremer, Jackson, & McPeek, 1968; Pomplum, Morgan, & Nellikunnel, 1992). Second, in many cases, test takers of different gender and ethnic groups seem to choose differently, which results in score biases to their disadvantage (Wainer & Thissen, 1992, 1993, 1994). To alleviate such biases, it has been advocated that scores

of differentially chosen CR items be compared and equated (Wainer, Wang, & Thissen, 1994). The equating theory, methodology, and some untestable assumptions have been investigated and explicated (see Thissen, Wainer, & Wang, 1994; Wainer & Thissen, 1992, 1993, 1994; Wang, Wainer, & Thissen, 1995).

The purpose of this study is to address some more fundamental issues—how test takers actually choose CR items on a test and how their choices impact their performances. The specific factors that affect test taker choices and have been investigated in this paper are

- test taker ability,
- curriculum,
- content familiarity, and
- types of choosing behavior, such as random, serial, and selective choosing methods.

It is through the investigation of the relationship among the above factors that a good understanding of how test takers choose items is achieved.

Research Instrument and Data Collection

Three types of data were collected and analyzed in this study. The first data set was the national data containing the responses and scores of 18,462 test takers from the College Board's 1989 Advanced Placement Chemistry Examination (The College Board, 1990). This test consisted of two sections: Section I contained 75 compulsory multiple-choice (MC) items that all test takers were expected to finish. Due to their high reliability and uniform standards, these MC items will be used as the principle measure of test takers' AP chemistry abilities. Section II had four parts (Part A, B, C, and D) of varying numbers of compulsory and optional essay items. The focus of this paper is on Part D of Section II. This part consisted of five essay-type questions from which test takers were instructed to choose three, forming ten choice combinations.

In terms of their content, four of the five essays can be classified as "core or general chemistry" subjects and one, as "a noncore or advanced" chemistry topic. Specifically, Essay 5 addresses valence and electric configuration; Essay 6, bonding principles; Essay 7, solution chemistry in the form of laboratory chemistry; and Essay 8, thermo-chemistry. The final, Essay 9, is concerned with the specialized topic of nuclear chemistry.

Table 1 shows that, of the 18,462 test takers who took the 1989 AP Chemistry Examination, 12,234 were male, and 6,228 were female. Seven major racial/ethnic groups were identified in these data. The test taker population was predominantly Caucasian (about 72%), while the second largest group was Asian (about 20%). The remaining groups made up about 4% of the test taker population. In most racial groups, there were almost twice as many male test takers as female test takers. Only in the group of African Americans were there more female test takers (319) than male test takers (271).

TABLE 1

Test taker composition of the 1989 Advanced Placement Chemistry Examination classified in terms of racial/ethnic groups and gender differences

Racial/Ethnic Groups	Female	Percent	Male	Percent	Total
Caucasian	3,964	23.13	8,323	48.57	12,287
Asian	1,262	7.36	2,173	12.68	3,435
African American	319	1.86	271	1.58	590
Latin American	97	0.57	196	1.14	293
Mexican	60	0.35	122	1.58	182
American Indian	20	0.12	26	0.15	46
Puerto Rican	16	0.09	27	0.16	43
Others	80	0.47	180	1.05	260
Total	5,818	33.95	11,318	66.05	17,136

Note. 1,326 test takers were missing because they did not identify their ethnicities.

The second data set was collected through an instrument called the "Advanced Placement Chemistry Survey and Test Kit" (the Kit). The Kit was administered to 554 chemistry students in the state of Hawaii whose ability distribution was shown to be comparable to the above national AP chemistry population. Detailed analyses and results on this data set are reported in a separate paper (Wang, 1996b). What is of direct relevance to this paper is Part D of the Kit from the second data set, which reprinted the same five essays mentioned previously, but with Essay 7 as the last essay to test the order effect. The Kit asked the Hawaii students to hypothetically choose three of the five essays. The specific question in the Kit was, "If you were asked to choose three of the five essays, which of the following combinations would you choose?" All the 10 three-essay combinations were presented to the students as a single 10-choice multiple-choice item. The purpose of this experiment was to find out whether or not the chemistry students in Hawaii could independently replicate the choice patterns of the national population.

In order to explain the choice patterns of both the national test taker population and the students of Hawaii, a survey called "Chemistry Teacher Expert Judgment Survey" (Appendix) was conducted with all AP chemistry teachers in the state of Hawaii regarding the five CR essay questions. This survey contained two main tasks. First, the teachers were asked to rank the five CR essays in terms of their difficulty. The purpose of this task was to see if the teachers' perceptions agreed with the popularity of essay choices found in the national data. Second, the teachers were asked to assess the effect of the AP chemistry curriculum on choices, to evaluate to what extent their AP chemistry curriculum covered the topics addressed by each of the five essay questions.

Analyses and Results

The analyses of this paper were carried out in four phases. Phase I concentrated on how test takers actually made choices in Part D of the AP chemistry exam and how those choices affected their performance. The choices made by the chemistry students of Hawaii were compared with those of the national population. Based on the survey data from the AP chemistry teachers in Hawaii, Phase II investigated why test takers chose their essays from the curricular and instructional perspectives. Phase III explored how test takers of various abilities chose, and whether or not there was any difference in performance between those test takers who made deliberate choices and those who made casual or random choices. Phase IV compared the performances of the test takers who seemed to have consciously chosen with those who did not. Various research questions will be presented and answered in this section of the paper.

Phase I: Overall Essay Choice Tendencies and Performance

How Did Test Takers Choose the Five Essays, Similarly or Differently?

Allowing test takers to choose three of the five essays yields 10 essay combinations ($_5C_3$): Essays 5, 6, and 7; Essays 5, 6, and 8; Essays 5, 6, and 9; Essays 5, 7, and 8; Essays 5, 7, and 9; Essays 5, 8, and 9; Essays 6, 7, and 8; Essays 6, 7, and 9; Essays 6, 8, and 9; and Essays 7, 8, and 9. The first interesting questions are "How did the test takers choose them? Did they choose them more or less equally?"

A look at Figure 1 reveals the five essays were chosen in drastically different ways. Essay Combination 5, 6, and 8 was the most popular one, while Essay Combination 6, 7, and 9 was the least popular. Over 55% of the total population (9,425 of the 18,462 test takers) chose the two essay combinations 5, 6, and 8 and 5, 7, and 8; while less than 1% of the test takers chose Essay Combination 6, 7, and 9. The rankings of the choice combinations are summarized in Table 2.

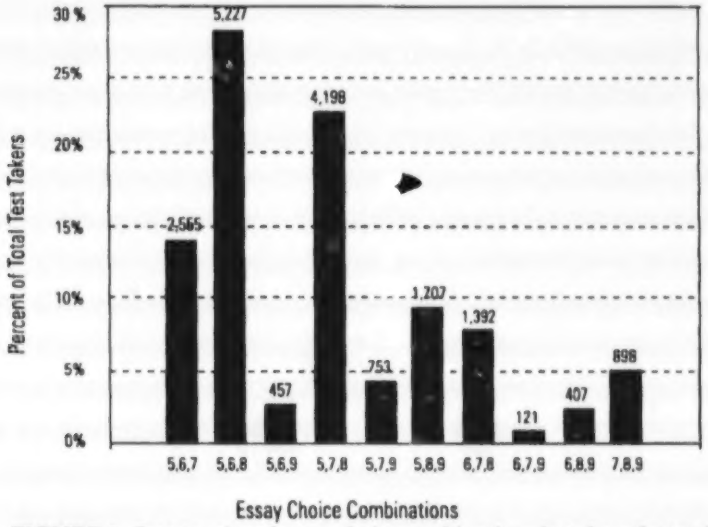

FIGURE 1. *Frequencies of essay choice combinations based on the total population of 18,435 test takers*

TABLE 2

Rank of choices of essay combinations and actual frequencies

	Rank										
	1	2	3	4	5	6	7	8	9	10	Total
Essay choice	5,6,8	5,7,8	5,6,7	5,8,9	6,7,8	7,8,9	5,7,9	5,6,9	6,8,9	6,7,9	10
Choice frequency	5,227	4,198	2,555	1,707	1,392	898	753	457	407	121	18,435
Percent	28.35	26.68	14.86	9.26	7.55	4.87	4.08	2.48	2.21	0.66	100%

The popularity rankings of the five individual essays are obtained by summing the number of times each of the five essays was chosen, and is displayed in Figure 2. Essay 5, on valence and electronic configuration, was the most popular, followed by Essay 6. Essay 9, on nuclear chemistry, was the least popular.

It is clear that the popularity rankings of the five essays correlates perfectly with the order in which these essays were presented on the test, an indication of sequence effect.

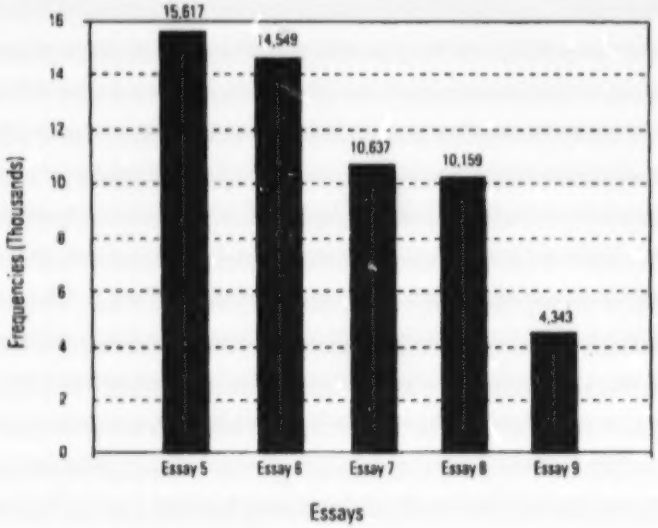

FIGURE 2. *Popularity rank of the five essays*

Are the Essays of Similar Difficulty?

Why did the test takers choose the essays so differently? In addition to the different essay contents and sequencing effects, another reason may have been that these essays were differentially difficult. Due to the lack of a specifically designated index for essay difficulty, two alternative measures are used as indicators of relative essay difficulty—the item response theory (IRT) expected scores and the observed raw scores for each of the five essays, conditioned on abilities.

The IRT expected score is obtained through a common-item anchorage. It is known that all the 10 essay combinations share at least one essay in a chained fashion. More specifically, Essay Combination 5, 6, and 7 is connected with Essay Combination 7, 8, and 9 through Essay 7, called the "anchor item." Through such anchoring, the probability curve of answering each of the graded responses is calculated through Bock's nominal model (1972). The expected scores are computed by aggregating across the trace lines through

$$E(Score \mid \theta) = \Sigma x_j T_{jx}(\theta) ,$$

where θ is the test taker's ability; m, the number of categories; and x, the category number.

Figure 3 displays the IRT expected scores of the five essays on the basis of 18,462 test takers in the 1989 AP Chemistry Examination. There is a substantial amount of interaction between the expected scores of the five essays and ability distributions. In general, Essays 5 and 8 are highly similar to each other in their expected scores across the entire ability distribution. Essay 9 is the second most difficult for the very low-ability range, but becomes the easiest for the middle-ability range onward. Essay 6 seems to be the most difficult of all the essays for the low-ability through the upper-middle-ability range and switches to be the second or third easiest for the high-ability test takers. Essay 7 parallels Essays 5 and 8 in its difficulty within the low- to upper-middle-ability range and then becomes the easiest for the extremely high-ability test takers.

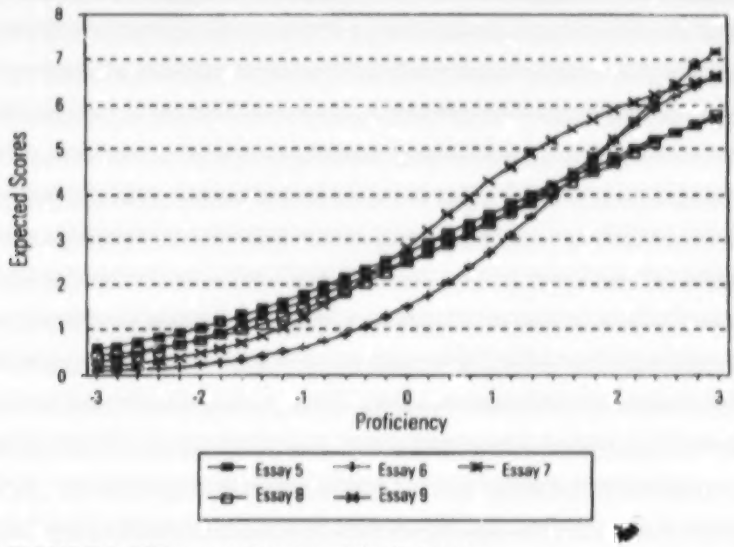

FIGURE 3. *IRT expected scores of the five essays*

Based on the total scores on the 75 multiple-choice questions, the 18,462 test takers were divided into 10 ability groups. Figure 4 shows the mean scores of the five essays for 10 ability levels. It can be seen that the mean score curves of the five essays parallel those of the expected scores of the five essays in Figure 3, except the conditional mean score curves indicate little interaction. Although Essay 9 remains the easiest, Essay 6 becomes the hardest essay instead of Essay 5, as shown earlier. Essays 7 and 8 are virtually identical in terms of their difficulty throughout the entire ability range. Although Essay 5 is close to Essays 7 and 8 for the lower half of the ability distribution, it becomes slightly easier for the upper half of the ability distribution. In their entirety, Essay 9 is the easiest, followed by Essays 5, 8, and 7, and ending with Essay 6 as the hardest. Table 3 summarizes the mean scores on the five essays.

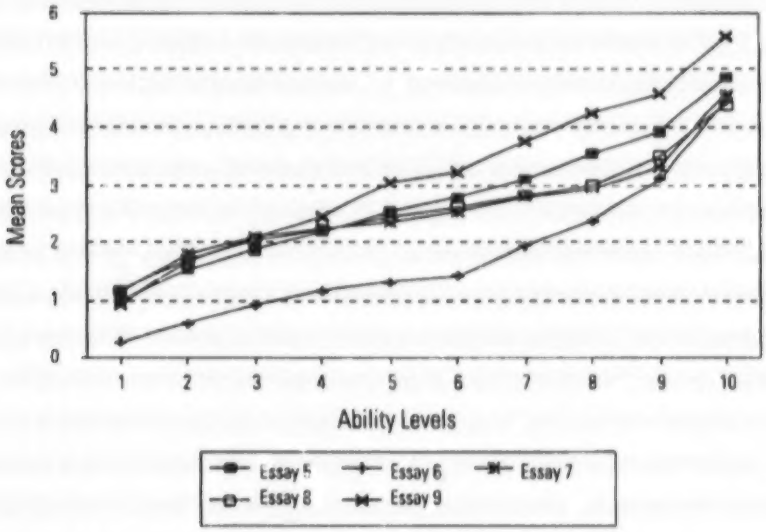

FIGURE 4. *Essay difficulty as reflected by conditional mean scores*

TABLE 3

Mean scores on five essays

	Essays				
	5	6	7	8	9
Maximum score	8	8	8	8	8
Mean score	2.71	1.61	2.50	2.65	3.41
Standard deviation	1.96	2.02	1.88	2.01	2.32
Mean as percent of maximum	34	20	31	33	43

The order of the five essays in terms of mean scores is 9, 5, 8, 7, and 6, with Essay 9 having the highest mean of 3.41 and Essay 6 having the lowest mean of 1.61 points. The means for Essays 7 and 8 are very similar. Compared with the maximum score of eight points possible on each essay, the mean scores ranging from 1.61 to 3.41 points, are fairly low, at only 20% to 40% of the maximum scores. Such averages imply that these essays were not only substantially difficult for the test takers as a whole, but also considerably different in terms of difficulty. The difference between the highest and lowest means is 1.8 points, occupying 23% of the maximum score.

It should be pointed out that the accuracy of either of the two indices is open to debate for two reasons: First, the IRT expected scores are obtained under the violation of one of the key assumptions of IRT missing-at-randomness because test takers deliberately chose their essays, answering some and omitting others. Second, the conditional observed mean scores might not be an unbiased choice either—they were based on the test takers who chose their own essay items and the difficulty levels of the essay items were not equated, as in IRT expected scores.

Did Test Takers Score Higher on More Frequently Chosen Items?

The answer appears to be "no" as summarized in Table 4, which presents the relationship between the ranking of essay choice preference and mean performance. The correlation between the popularity ranking of the five essays and their corresponding means is -0.60. The correlation between the ranking of essay combinations and their mean scores is -0.22. Regardless of their statistical significance, these negative correlations are somewhat surprising because higher scores are expected for the more popular essay combinations, since it is logical to assume that test takers would choose the items that they feel confident in.

TABLE 4

Relationship between rank of essay choice preference and mean performance

	Rank									
	1	2	3	4	5	6	7	8	9	10
Essay choice	5,6,8	5,7,8	5,6,7	5,8,9	6,7,8	7,8,9	5,7,9	5,6,9	6,8,9	6,7,9
Choice frequency	5,227	4,198	2,555	1,707	1,392	898	753	457	407	121
Mean scores	7.10	8.14	4.57	9.89	7.18	9.72	8.37	7.85	8.75	7.12

Did Allowing Choice Equally Affect the Performance of Test Takers of Various Abilities?

Again, the answer appears to be "no." According to Figure 5, these five essays are differentially difficult to test takers of similar abilities. Two general trends can be observed. First, as expected, the higher the ability levels, the better the respondents perform across all five essays. The mean difference between level-1 and level-10 test takers is 3.88.

Second, the effect of choice has a bigger impact on the test takers of middle-ability levels than those of extreme-ability levels. For example, the maximum differences resulting from choosing among the five essays are 0.89 points (1.15 - 0.26) for level-1 test takers as compared to 1.18 points (5.56 - 4.38) for level-10 test takers. Yet, the differences from choosing among these five essays range from 1.72 points (3.03 - 1.31) for level-5 test takers to 1.83 points (3.76 - 1.93) for level-7 test takers. Although small in its magnitude, a difference of 1.83 points is 22% of the maximum score of 8 points.

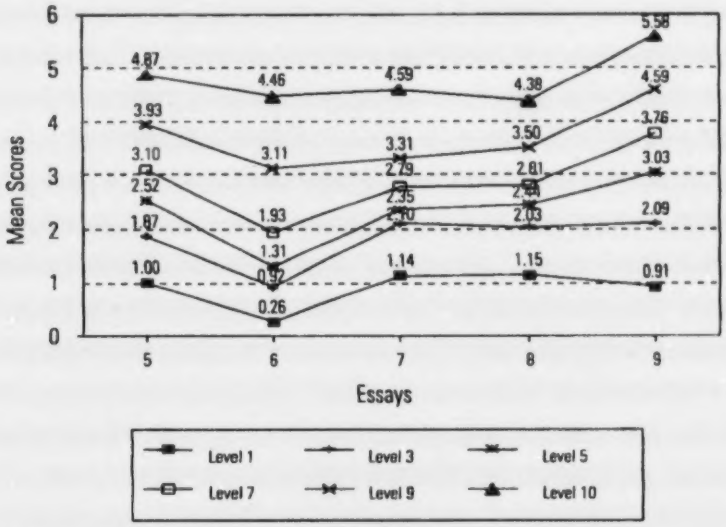

FIGURE 5. *Mean scores of five essays conditioned on ability levels*

The above findings seem to suggest that the scores of middle-level test takers are affected more than the scores of either extremely high- or low-ability test takers. A little thought will find such a phenomenon justifiable. If test takers are of either extremely high or low ability, it is highly likely that they would either get an item right or wrong, respectively. Yet, if they are of middle ability, the choice of an item that is just "right" for their level could make a large difference in how they would perform. If they chose a right item, they would perform well. If the chosen item is above their level, they might perform poorly.

What do the mean scores of the 10 three-essay combinations look like at different ability levels? Figure 6 displays similar patterns of mean scores as those observed earlier. Again, there is less variation in the mean scores at either very high- or very low-ability levels, while the mean scores of middle-ability test takers are affected a great deal more.

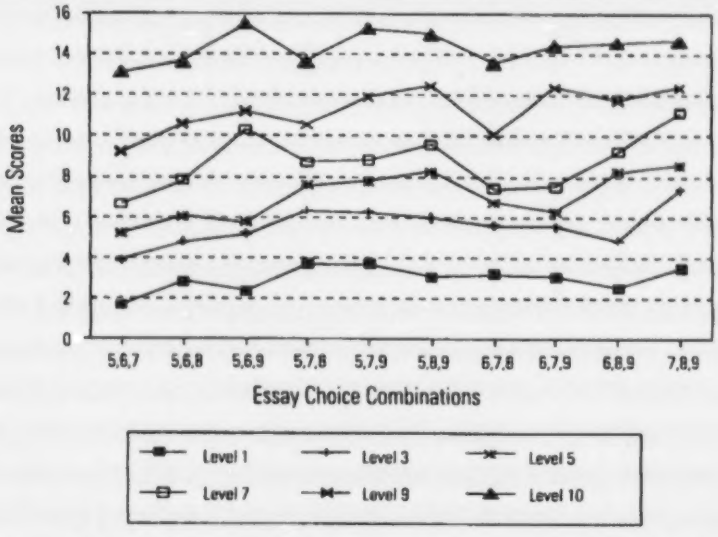

FIGURE 6. *Mean scores on 10 essay combinations conditioned on ability levels*

Would Chemistry Students in Hawaii Choose the Five Essays in a Similar Way as Their National Counterparts?

The answer appears to be "yes." It can be seen from Figure 7 that the overall choice pattern for the 10 essay combinations made by the 554 Hawaii students who responded to the essay choice question substantially mirrors the choice pattern made by the 18,462 test takers of the 1989 AP Chemistry Exam. The correlation between the two patterns is .87.

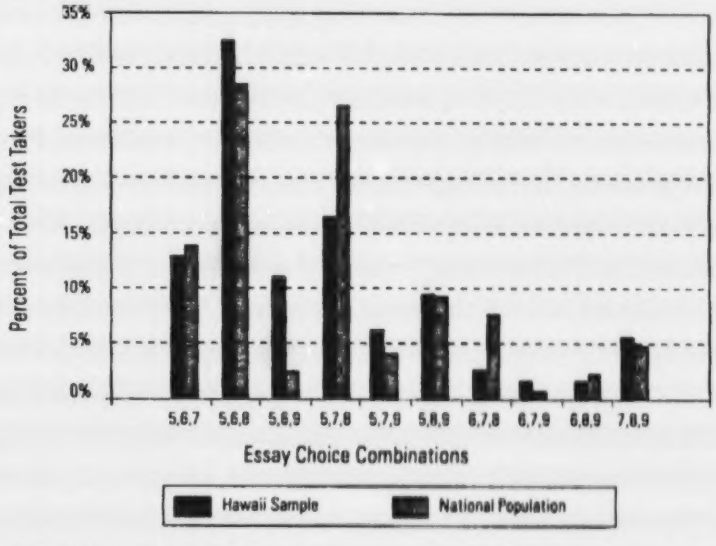

FIGURE 7. *Comparison of the overall essay choice pattern of 554 Hawaii and 18,462 national test takers*

Three features are worth pointing out. First, like the 1989 AP Chemistry Exam population, the Hawaii students also liked Essay Combination 5, 6, and 8 the best. Essay Combination 5, 7, and 8 still remains the second most popular combination, although the absolute number of test takers decreased considerably. Second, almost the same proportion of Hawaii students (13%) as the 1989 AP test takers (14%) chose Essay Combination 5, 6, and 7. Note that Essay 7 was presented as the last essay in the Kit instead of Essay 9, in order to test the effect of positioning on Essay 9. There seems to be something inherent about Essay Combination 5, 6, and 7 that attracts some test takers. Third, probably due to the changed position of Essay 9, there seems to be an increase in the choice frequency for Essay 9 by Hawaii students. For example, only about 3% of the 1989 AP test takers chose Essay Combination 5, 6, and 9, while about 11% of the Hawaii students chose it. There are also more Hawaii students choosing Essay Combinations 5, 7, and 9 and 7, 8, and 9.

Phase II: Curricular Explanations of Differential Essay Choices

How could Hawaii students, more or less, replicate the seemingly divergent essay choices of the national population? This mystery was accounted for by surveying eight AP chemistry teachers in Hawaii. As shown in the Appendix, the "Chemistry Teacher Expert Judgment Survey" asked the teachers to perform three tasks: (1) to rank the five essays by difficulty; (2) to estimate AP chemistry textbook coverage of the chemistry topics represented by the five essays; and (3) to elaborate on the common characteristics shared by various essay combinations. The purpose of this survey is three-fold: (1) to find out whether or not the teachers' rankings agreed with test takers' choice preferences; (2) to determine the effects of textbook coverage on test takers' choice preferences; and (3) to gain expert explanations on the differential choices. The results on the first two tasks are summarized in Table 5.

TABLE 5

Summary of teachers' rankings of essay difficulty levels

	Difficulty Ranking					Extent of Textbook Coverage				
Teachers	5	6	7	8	9	5	6	7	8	9
1	2	5	1	4	3	5	5	4	5	3
2	1	4	3	4	5	2	2	2	4	1
3	1	3	2	4	5	5	4	3	2	1
4	1	2	4	3	5	3	4	2	4	3
5	2	1	5	3	4	3	2	1	3	2
6	1	3	5	2	4	5	5	3	5	4
7	1	.	.	.	.	4	3	3	4	1
8	.	1	.	.	.	3	3	3	3	3
Average	1.4	3	3.3	3.3	4.3	3.75	3.5	2.6	3.7	2.3

The teachers tended to rank the core-chemistry items easier than the noncore-chemistry item. Specifically, the eight teachers, on average, ranked Essay 5 as the easiest, Essays 6, 7, and 8 as approximately the same, and Essay 9 as the most difficult. Such rankings agreed well with the popularity of the five essays with a correlation of 0.92, but completely disagrees with the empirical mean scores of the five essays with a -1.0 correlation. The national test takers scored higher on the essays that the teachers deemed more difficult. Furthermore, the teachers' rankings of essay difficulty correlated at -0.73 with their ratings of the textbook coverage of the five essays, which signifies that the teachers tended to rate less covered chemistry topics more difficult.

What can be inferred from the above finding? It seems to suggest that the teachers could not accurately evaluate the difficulty of the essays, and that they tend to think the more commonly taught and familiar chemistry subjects are easier than less taught ones. If AP chemistry students can be assumed to perceive the five essays in the same way as the teachers, which is highly likely because the former learn from the latter, then the negative correlation between essay difficulty and choice popularity is no longer surprising. The test takers would tend to equate content familiarity with essay difficulty, believing the more familiar items to be

the easier ones. (In Phase III of the analyses of this paper, it will be shown that test takers of all ability levels chose virtually the same way.) The following are the teachers' explanations on some of the essay combinations. Note that all the words in quotation marks are direct quotes from the teachers' written comments.

Why Was Essay Combination 5, 6, and 8 the Most Popular Essay Combination?

Why did 29% of the national test taker population choose Essay Combination 5, 6, and 8? There seem to be two reasons. First, contentwise, these three questions address the core-chemistry theories and principles that are covered in regular high school chemistry, and re-emphasized in AP chemistry textbooks. Second, taskwise, the questions involve accounting for phenomena and describing concepts, which seem to be less complicated than actually solving a problem as required by Essay 7. Students should feel "safer" to "talk about" something than actually implementing something. Students, especially low-ability ones, might have felt they could, at least, obtain partial credit by describing the chemistry phenomena in question. This combination "is also the best place to avoid lab experiences as required by Essay 7 or the much less taught subject of nuclear chemistry of Essay 9."

Why does the most popular combination have the second lowest mean score? Although familiar to students, the subquestions or parts of the three essays, especially Essays 6 and 8, are very challenging because they require deep understanding of the system within which these questions are framed, such as "the expected trends in the melting points of the four compounds LiF, NaCl, KBr, and CsI." Such knowledge is so subtle that it "often escapes most students the first time around."

Why Did Essay Combination 5, 6, and 7 Have the Lowest Mean Score?

Described as a "nightmare" by one teacher, Essay Combination 5, 6, and 7 includes not only the theoretically challenging Essay 6, but also the laboratory-experience-demanding Essay 7, although Essay 5 is relatively easy. "Many test takers might not have realized how tricky this could be, since this required not only deep theoretical understanding and good computational skills, but also extensive laboratory experiences, which many students lack."

Why Was Essay Combination 6, 7, and 9 the Least Popular Combination?

Given the above explanations on Essay 6 and Essay 7, and with Essay 9 addressing such a highly specialized subject as nuclear chemistry (see more detailed reasons below), it is no longer difficult to explain why Essay Combination 6, 7, and 9 had the lowest choice frequency.

Why Did Essay 9 Have the Highest Mean Score but the Lowest Choice Frequency?

The reason that Essay 9 was the least popular among all the essays is that nuclear chemistry is often discussed towards the end of most AP textbooks. Two examples of the most widely used AP chemistry textbooks are *Chemical Principles* by Masterton, Slowinski, and Stanitski (1985) and *Chemistry* by R. Chang (1988), in which nuclear chemistry is one of the last chapters, for example, the 23rd chapter or later. "Very often, the subject of nuclear chemistry is either not reached or glossed over by many chemistry teachers." As a result, it is highly unlikely that the insufficient instructional coverage would have generated much test taker enthusiasm for choosing it.

The reason Essay 9 was the easiest is that it appears to be testing only the recall of particular facts on the decay of a particular isotope, rather than extensive interconnected chemistry knowledge as the other essays do. It is mostly an either "you know" or "you don't know" situation. If you had studied it before, you had a high chance of answering it correctly. But if you had not studied this particular problem, it was virtually impossible to score any points. As a result, those who did not know anything about it would not attempt it. Those who did know about it would score high. This explains why this essay had the highest mean score in spite of having the lowest choice frequency.

Why Did Essay Combination 5, 8, and 9 Have the Highest Score?

In light of the previous descriptions, three reasons can be deduced. First, these three essays share one common feature—they are detail- and fact-oriented. In order to successfully complete the theoretical and conceptual explanation, the students who chose this combination had to have an exceptionally thorough knowledge of chemistry. Second, because of the relatively late location of nuclear chemistry in AP chemistry textbooks, a student who has reached and studied the chapters on nuclear chemistry, to a certain extent, should be already well educated in chemistry. It will be shown later in this paper that more higher ability than lower ability test takers chose this essay combination. Finally, the noninclusion of Essays 6 and 7, which required deep theorization and extensive laboratory experiences, respectively, must have slightly eased the overall complexity for this combination.

Phase III: Relationship Among Test Taker Essay Choices, Ability Profiles, and Performance

With the above curricular and instructional explanations to the perplexing relationships between essay choices and performance found in Phase I, the next step is to find out (1) who chose the various essay combinations and, especially, (2) who accurately identified the relative difficulties of the essays and chose the easier ones to their advantage. These two questions are the center of investigation in Phase III.

What Was the Test Taker Ability Distribution Behind Each Essay Choice Combination?

The test taker ability distribution behind each essay choice combination is displayed in Figure 8. Figure 8 consists of 10 plates, each depicting the number of test takers for each of the 10 ability levels who chose each of the 10 essay combinations. It can be seen that, except for Essay Combination 5, 6, and 7 and Essay Combination 5, 8, and 9, the same essay combinations were chosen, more or less, by all ability levels—although the actual number of test takers differed. Essay Combination 5, 6, and 7 and Essay Combination 5, 8, and 9 had opposite ability distributions. The former had more low-ability test takers, while the latter, more high-ability test takers. More results will be reported on these two combinations later. It can be concluded that, generally, both low- and high-ability students tended to perceive and choose similarly.

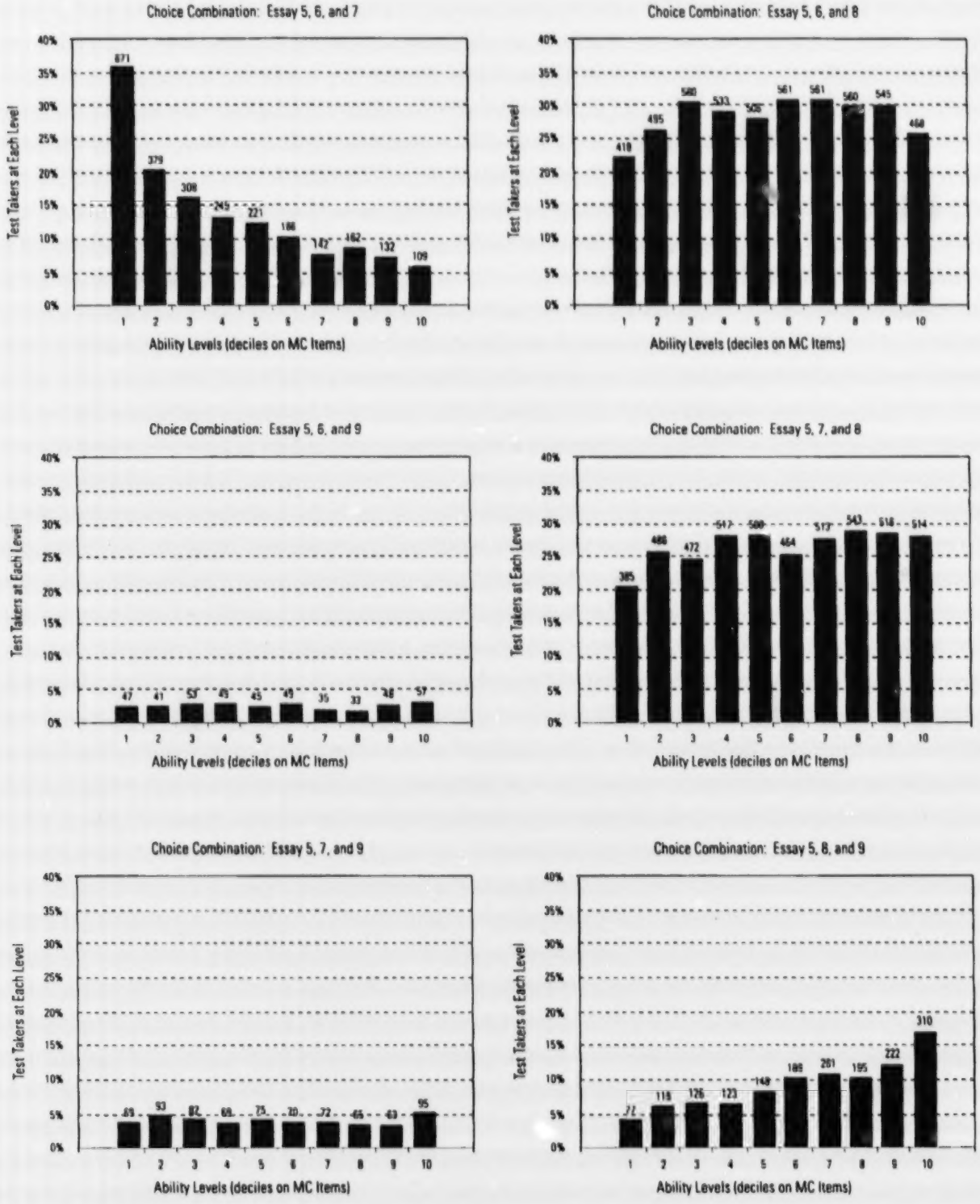
Choice Combination: Essay 5, 6, and 7
Test Takers at Each Level
Ability Levels (deciles on MC Items)
Choice Combination: Essay 5, 6, and 8
Test Takers at Each Level
Ability Levels (deciles on MC Items)
Choice Combination: Essay 5, 6, and 9
Test Takers at Each Level
Ability Levels (deciles on MC Items)
Choice Combination: Essay 5, 7, and 8
Test Takers at Each Level
Ability Levels (deciles on MC Items)
Choice Combination: Essay 5, 7, and 9
Test Takers at Each Level
Ability Levels (deciles on MC Items)
Choice Combination: Essay 5, 8, and 9
Test Takers at Each Level
Ability Levels (deciles on MC Items)

FIGURE 8. *Test taker ability profiles behind each choice pattern*

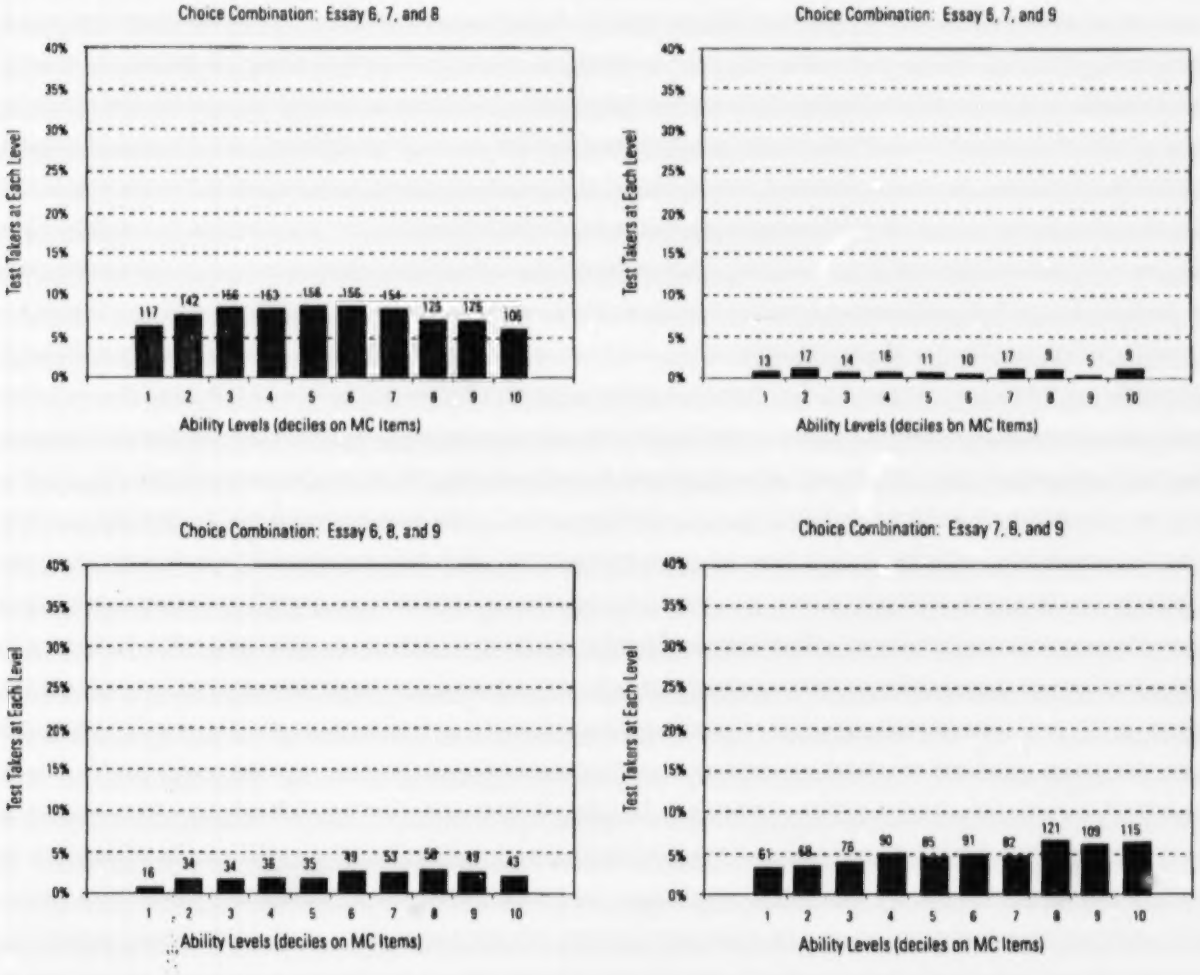

FIGURE 8. (continued) *Test taker ability profiles behind each choice pattern*

How Did Test Takers of the 10 Ability Levels Choose, Respectively?

Complimentary to the previous investigation of "who chose what" is another look into the essay choice patterns across the 10 ability levels. Did test takers of different abilities have similar or different essay choice patterns? Figure 9 consists of 10 graphs, each of which represents the choice pattern of each of the 10 levels of test takers. It can be seen that test takers from level 2 through level 10 tended to choose the 10 essay combinations similarly. The extent of the similarity of choice patterns is reflected by the significant intercorrelations among the choice frequencies of the 10 ability levels summarized in Table 6.

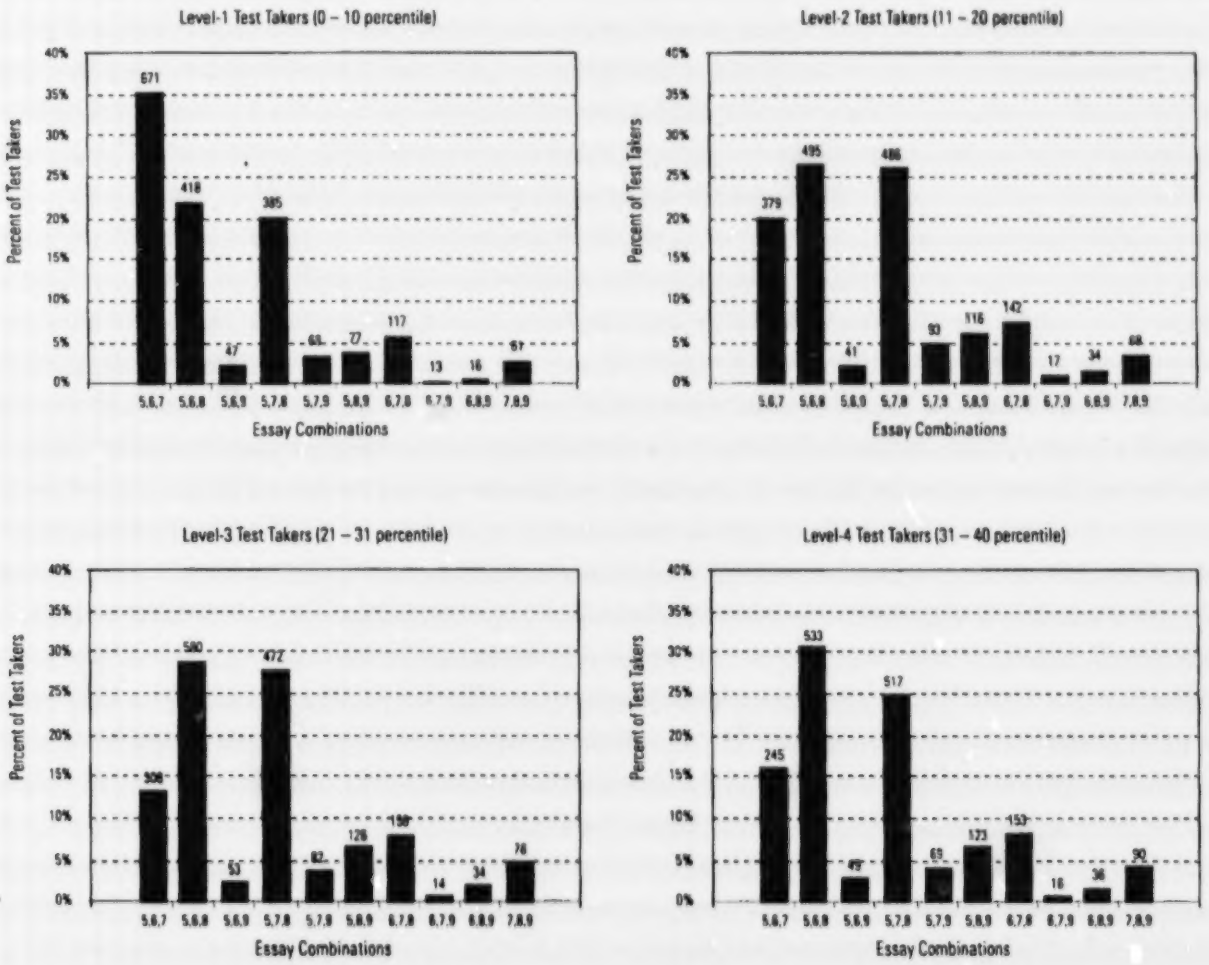

FIGURE 9. *Essay choice patterns of test takers of 10 ability levels*

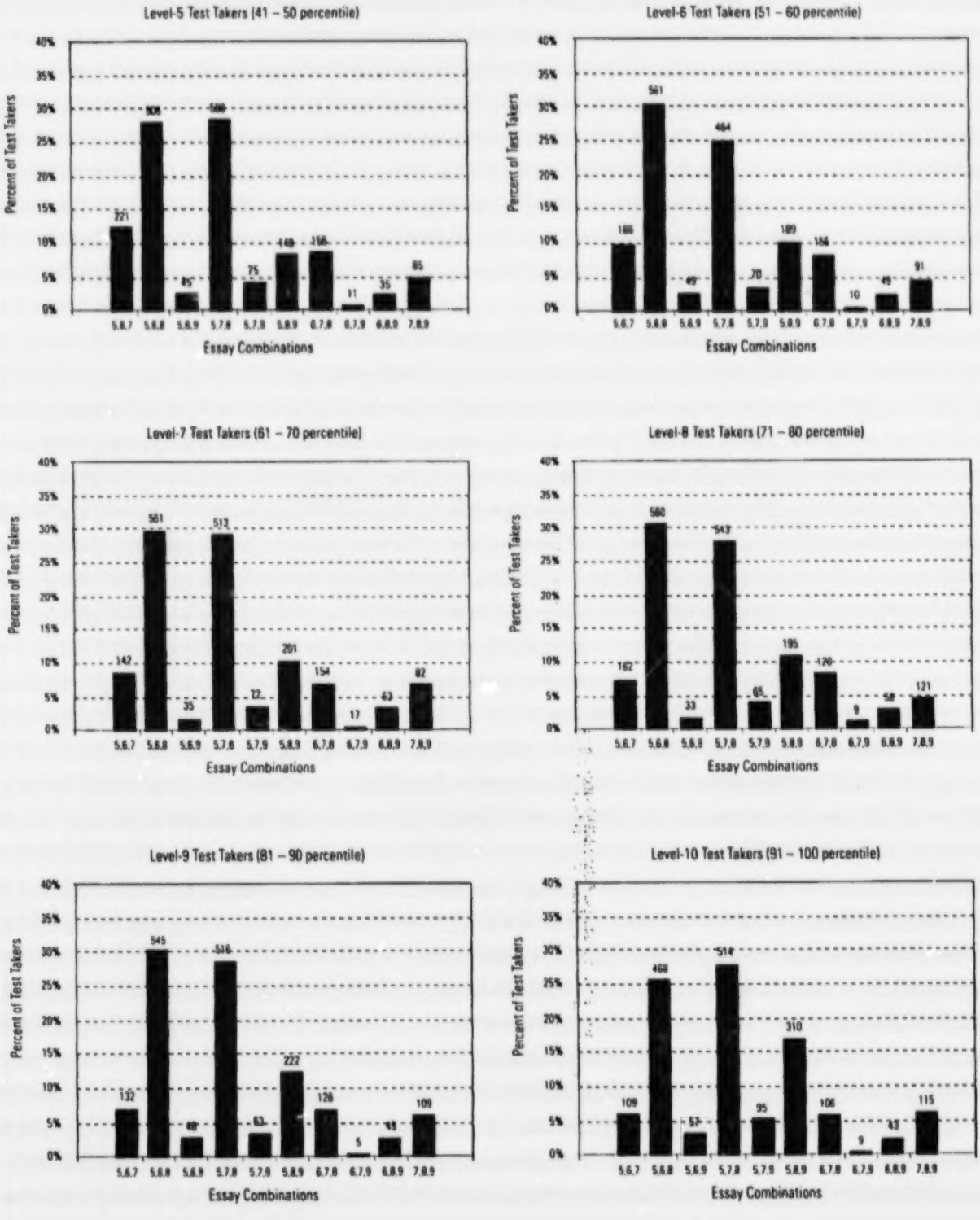

FIGURE 9. (continued) *Essay choice patterns of test takers of 10 ability levels*

BEST COPY AVAILABLE

TABLE 6

Intercorrelations among choice preferences of test takers of 10 ability levels

	Ability Levels										
Ability Levels	1	2	3	4	5	6	7	8	9	10	All
1	1.00										
2	0.88	1.00									
3	0.79	0.98	1.00								
4	0.73	0.97	0.99	1.00							
5	0.71	0.96	0.98	1.00	1.00						
6	0.65	0.92	0.97	0.98	0.99	1.00					
7	0.58	0.90	0.95	0.97	0.98	0.99	1.00				
8	0.61	0.91	0.95	0.98	0.99	0.99	1.00	1.00			
9	0.57	0.88	0.93	0.96	0.98	0.99	1.00	1.00	1.00		
10	0.49	0.81	0.86	0.90	0.93	0.94	0.96	0.96	0.98	1.00	
All	0.74	0.96	0.99	0.99	0.99	0.99	0.97	0.98	0.97	0.92	1.00

Note. Correlations below 0.65 are not statistically significant at 0.05 level.

Two trends can be observed from Figure 9 and Table 6. First, the choice pattern of level-1 test takers is significantly different from those of level-7 through level-10 test takers. Second, the choice patterns of test takers of levels 2 to 4 are highly similar, while those of test takers above level 5 are virtually identical. It can be concluded that there is a substantial amount of agreement regarding the choice patterns of the test takers for all ability levels, except for level-1 test takers. The biggest variation occurred with level-1 test takers.

The above finding is contrary to what has been commonly hypothesized. A lot more variations in the choice patterns are expected across the 10 levels of test takers, due to the assumption that high- or higher-ability test takers would choose differently from low- or lower-ability test takers. What has been found is that equal numbers of test takers from ability levels 2 to 10 balked at answering Essay 9 on nuclear chemistry. Equal numbers of test takers from the 10 ability levels consciously stayed away from the challenging Essays 6 and 7 simultaneously. Such a finding seems to suggest that the majority of the test takers across the 10 ability levels similarly perceived the nature and properties of all the five essays.

Who Chose the Essay Combination 5, 8, and 9, Which Yielded the Highest Mean Score?

Given the similar test taker ability profiles behind the 10 essay combinations, one would wonder who was really "bold" enough to confront the "risky" combination of 5, 8, and 9. Figure 10 (as reproduced from the sixth graph of Figure 8) shows that increasing numbers of higher-ability test takers chose this combination. This kind of "step-up" ability distribution seems to suggest two interesting phenomena. First, if this essay combination represents a relatively difficult one because of the infrequently taught subject of Essay 9, the "step-up" ability distribution seems to imply that some of the higher-ability test takers tended to choose a more challenging set of essays. On the other hand, if this essay combination represents relatively straightforward essays (these three essays have been described as the least "involved" set), then this step-up ability distribution may imply that some higher ability is required by test takers to accurately identify the real nature of the five essays.

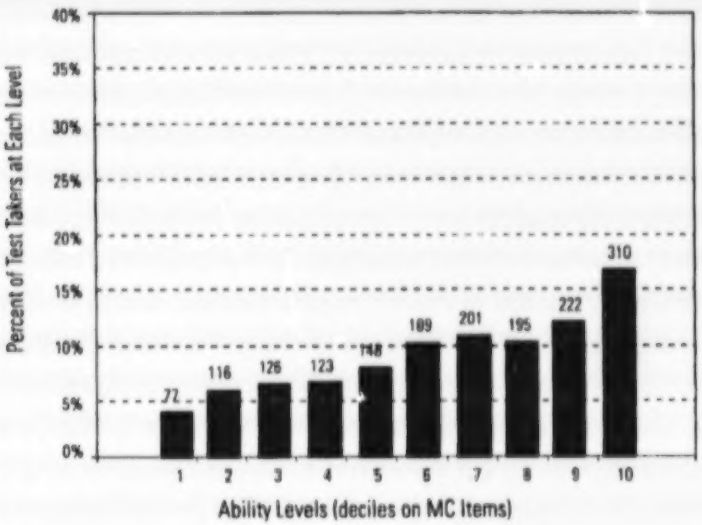

FIGURE 10. *Test taker ability composition in essay combination 5, 8, and 9*

Who Chose the Essay Combination 5, 6, and 7, Which Yielded the Highest Mean Score?

Who were really "ambitious" enough to take the "heavy-duty" combinations of 5, 6, and 7? Figure 11 (reproduced from the first graph of Figure 8) portrays a completely opposite picture of test taker ability profiles as compared to Figure 10. With the lowest mean score, Essay Combination 5, 6, and 7 has the highest number of level-1 test takers and the smallest number of level-10 test takers. As the ability level increases, the numbers of test takers steadily decreases. Such a "step-down" ability distribution seems to suggest that low-ability test takers did not seem to recognize the relatively high difficulty levels of this essay combination.

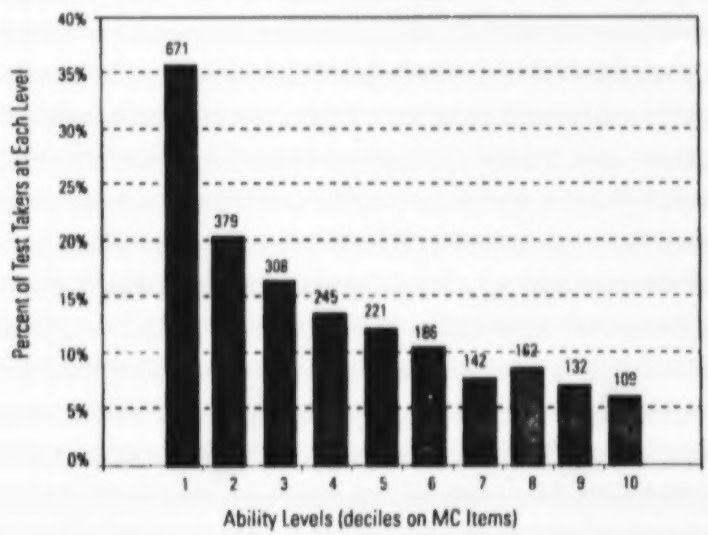

FIGURE 11. *Test taker ability composition in essay combination 5, 6, and 7*

Why did so many level-1 test takers choose this combination, given that Essays 6 and 7 are so complicated? Given the fact that a disproportionate 35% of level-1 test takers vs. only 6% of level-10 test takers chose this pattern, and that Essays 6 and 7 are so difficult, one can't help wondering how many of these level-1 test takers *really* made their choices deliberately.

Phase IV: Exploring the Effect of Sequential vs. Selective Choosing on Performance

Is it possible that a certain number of test takers might have just attempted Essays 5, 6, and 7 on a "first-come, first-tackle" basis, since they happened to be the first three essays? To what extent did their scores suffer as a result of their choosing sequentially? Although no one can be perfectly sure who deliberately chose and who did not, it seems reasonable to assume that most of the nondeliberately choosing test takers must have been those that answered Essay Combination 5, 6, and 7 in sequential order. Those test takers who did not choose Essays 5, 6, and 7 in sequential order, as well as those who chose the other essay combinations can be classified as selectively choosing test takers.

Two points are worth emphasizing, first, the order in which Essays 5, 6, and 7 are chosen is essential here. If one test taker first worked on Essay 5, then Essay 7, and finally Essay 6, that person is classified as a selectively choosing test taker because he/she did make some conscious decision in deciding which essay problem should be tackled first, even though his/her essay choice combination is Essays 5, 6, and 7. Second, not all the test takers who chose to answer Essays 5, 6, and 7 sequentially are necessarily nonchoosing test takers, and some selectively choosing test takers might have deliberately chosen to answer the three essays in the sequential order. There is no information in the data that can help to identify the latter test takers.

The Effect of Sequential vs. Selective Choosing on Level-1 Test Taker Performance of Essay Combination 5, 6, and 7

Two trends are obvious from Figure 12, which displays the numbers of sequentially versus selectively choosing test takers for various scoring points. First, although there were almost equal numbers of sequentially choosing and selectively choosing test takers (330 vs. 344) at ability level 1, seven times more sequentially choosing test takers received zero scores than selectively choosing test takers (210 vs. 30). Second, there is a significant mean score difference between the sequentially choosing and selectively choosing level-1 test takers. The former has a mean of 0.60 of 24 points, while the latter, a mean of 2.7. A t-test of the difference yields a t-value of 18.23, highly statistically significant at .0001 probability level.

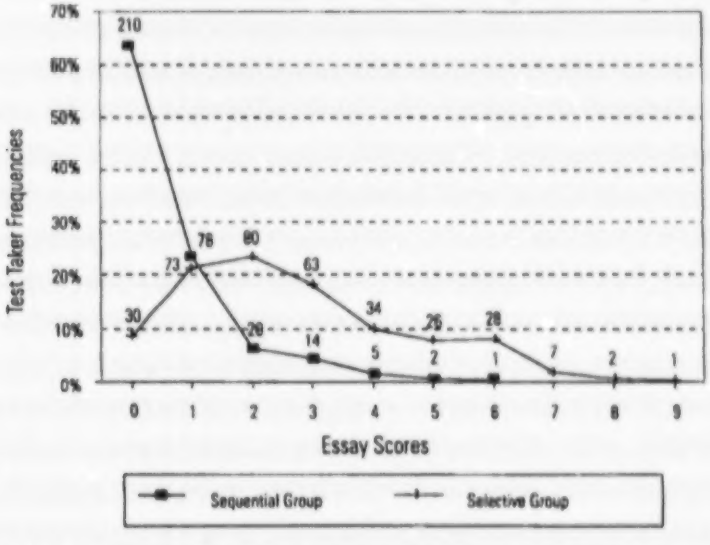

FIGURE 12. *Essay combination 5, 6, and 7 performance comparison between sequential vs. selective choosing of level-1 test takers only*

Why did the sequentially choosing test takers perform so poorly compared with selectively choosing test takers on Essay Combination 5, 6, and 7? Is it because the sequential group is much less competent in chemistry than the selective group? Not necessarily. The mean scores for these two groups differed by only 1.45 on Section I of the 75 MC questions; the sequential group had a mean of 4.25, while the selective group, 5.70. No one would attribute the dramatic difference in performance observed in Figure 12 to this minute mean difference. It is the belief of both this author and some teachers in Hawaii that a combination of factors might have attributed to the drastic differences between sequential choosing and selective choosing, among which the main factors were physical fatigue, mental stress, boredom, and so on. However, the most significant factor might have been personal motivation or lethargy. This hypothesis will be further strengthened by the following investigation of the performance differences between the two groups across the 10 ability levels.

The Effect of Sequential vs. Selective Choosing on the Performance Across 10 Ability Levels

Is it true that the difference between sequential and selective choosing only occurred for level-1 test takers? The answer is "no." Figure 13 shows the number of test takers who chose sequentially vs. selectively across the 10 ability levels. Two tendencies can be clearly observed, first, as the ability level increased, proportionally greater numbers of test takers chose selectively. Of the 2,566 test takers who responded to Essay Combination 5, 6, and 7, 1,856 test takers deliberately chose to answer these three essays, while approximately 710 test takers seemed to have answered these three essays because they were presented first. Second, as the ability increases, the number of sequentially choosing test takers drastically decreases. These two findings seem to suggest that test takers' ability is one of the major factors in determining whether or not they would choose sequentially. The higher the ability is, the more likely a test taker would choose selectively.

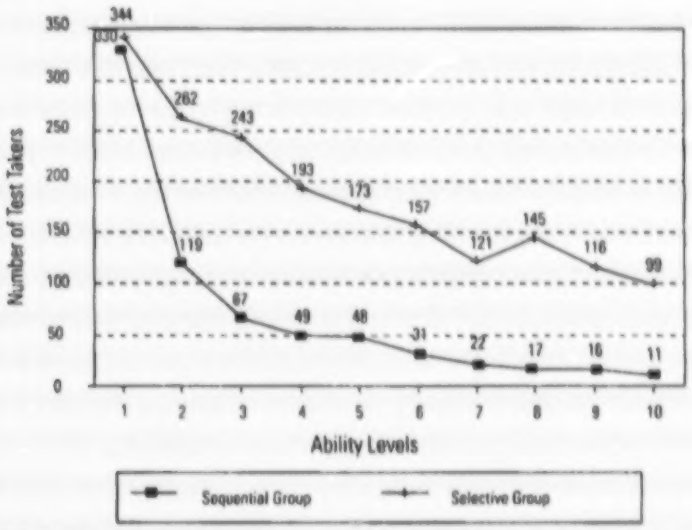

FIGURE 13. *Numbers of sequentially vs. selectively choosing test takers across 10 ability levels on essay combination 5, 6, and 7*

Is there a consistent difference in performance between these two groups on Essay Combination 5, 6, and 7? The answer is "yes," as shown in Figure 14. There is a consistent mean difference between the sequentially and selectively choosing test takers, ranging from the minimum of 2.06 points at level 6 to the maximum of 4.77 points at level 10. On the average, the selective group scored about 3.95 points better than the sequential group (5.66 for the selective group vs. 1.71 for the sequential group). Such differences are shown to be statistically significant by a two-way ANOVA, as shown in Table 7. The reason that the equal sample size of 11 randomly selected test takers in each cell was used was the drastically different numbers of test takers in each group—the biggest group size is 344, while the smallest group size is only 11. Since extremely different *N* sizes may ordinarily render ANOVAs invalid (Keppel, 1982), the author employs one of the

commonly used solutions to rectify the problem—random sampling of 11 cases from each of the 20 groups to make the sample sizes equal. Please note that the reduction of sample sizes makes the ANOVA much more conservative—making it more difficult to obtain significant results. Comparing Figure 14 with Figure 15 (which is a recalculation of Figure 14 using only the randomly sampled test takers) verifies that the random reduction of group sizes has essentially maintained all original performance characteristics of both the sequentially and selectively choosing test takers.

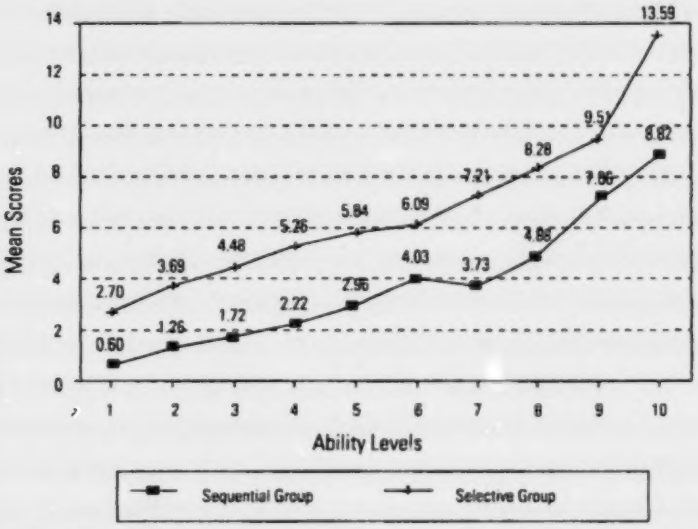

FIGURE 14. *Essay combination 5, 6, and 7 performance comparison between sequential vs. selective choosing across 10 ability levels*

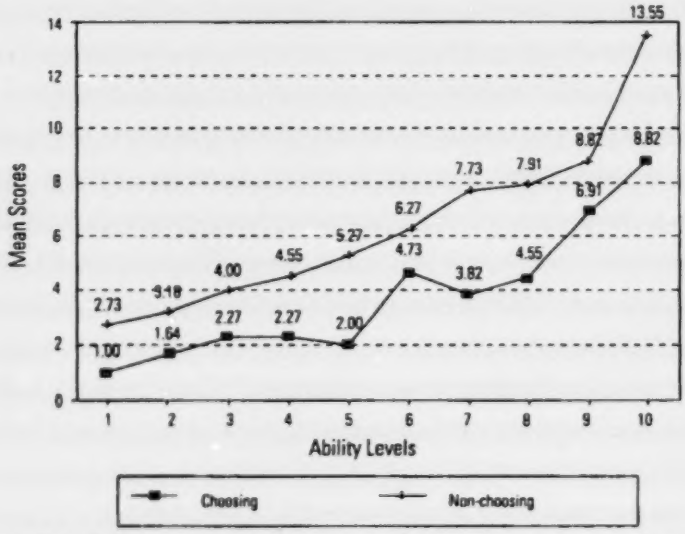

FIGURE 15. *Essay combination 5, 6, and 7 performance comparison between sequential vs. selective choosing*

TABLE 7

Summary of ANOVA results based on 11 randomly selected test takers across the sequential and selective groups and 10 ability levels

	DF	ANOVA SS	Mean Square	F Value	Prob >F	Effect Size
Model	19	2,043.44	107.55	10.64	0.0001	45.42%
Group	1	371.80	371.80	36.77	0.0001	8.87%
Level	9	1,607.80	178.64	17.67	0.0001	37.21%
Group x level interaction	9	63.84	7.09	0.70	0.7020	Negligible
Error	200	2,022.36	10.11			
Total	219	4,065.80				
R^2	0.503					

It should be pointed out that the observed consistent difference is somewhat surprising. Although differences between the two groups are expected at lower levels, they should gradually taper off as ability levels increase, because test takers with higher ability tend to solve problems more consistently. In both Figures 14 and 15, however, the difference between the two groups of test takers is the biggest at level 10. Such a finding seems to support the hypothesis made earlier that some test takers chose sequentially due to the factor of lethargy or motivation. Remember that Part D is the last part of a three-hour "grueling" AP chemistry examination. There is no doubt that a substantial amount of mental vigilance and effort is required to think about what it takes to solve an essay problem and to determine which essays are the most appropriate. Yet, it is likely that a considerable number of the sequentially choosing test takers must have said to themselves, "I have been working on the test too long and I am too tired to read through all of the five essays. Let me just get it done with and go home."

Do the above differences exist for any other essay combination? The answer is "no," since none of the other nine essay combinations involves the first three essays in Part D of the 1989 AP Chemistry Exam. Let's take Essay Combination 5, 6, and 8 as an example. Essay 8 is the fourth essay and choosing it involves a deliberate selection process on the part of the test taker. As Figure 16 shows, there is virtually no difference between the test takers who sequentially and selectively chose Essays 5, 6, and 8.

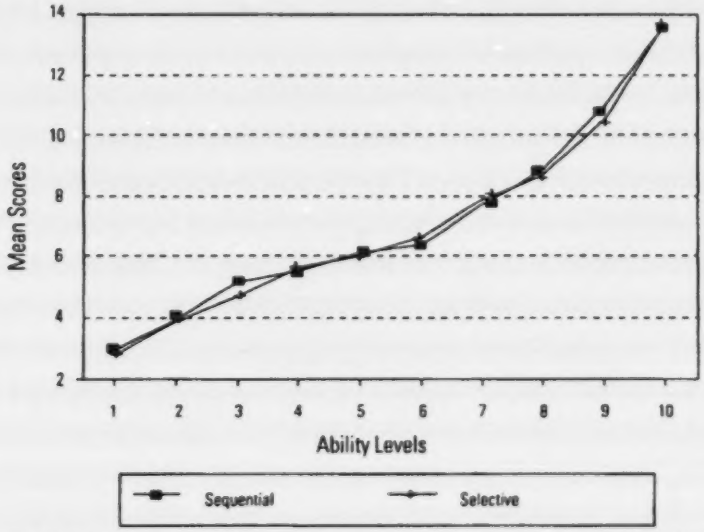

FIGURE 16. *Mean performance comparison between sequential and selective choosing on essay combination 5, 6, and 8*

Conclusions and Discussions

Through systematic analyses of the 1989 National AP chemistry data and the survey data on AP chemistry test takers and teachers in Hawaii, this study reported findings on nine areas of investigation concerning the relationships among the characteristics of CR items and test takers' choosing tendencies, ability levels, and test performance. Emphasis has been made not only on describing how test takers chose from a statistical point-of-view, but also on explaining why test takers chose the way they did, and the consequences of choices from both curricular and instructional perspectives. Such findings offer very useful suggestions and directions on how or whether choices should be incorporated into a computerized LSAT.

The first implication from this study relates to the need to control for the comparability of item difficulty and content for the items to be chosen, because it has been found that the source of divergent choices and consequently biases are due to the dissimilar essay items. If strict matching is not possible, a computer item selection algorithm could be used to select items with similar item characteristics.

When pretesting CR items is not possible, it has been shown by Wang (1996a, 1996b) that it is quite possible and simple to discern the similarities or dissimilarities of such items through multidimensional scaling methodologies without preexposing them to a large number of test takers to minimize test security breaches. One possible method to ensure such comparability is to offer test takers several similarly paired sets of CR items.

Based on findings of significant differences between test takers spanning the entire ability distribution who choose selectively or sequentially, it may be necessary to explicitly instruct test takers to read the CR items carefully and deliberately select them. In this way, test takers will be more likely to choose the most appropriate items for themselves and maximize their performance.

With the current trends toward holistic and performance assessment in either the traditional paper-and-pencil format or the more modern computerized adaptive mode, it is more and more frequent for CR items to appear on a test. In order to maintain the equity of scores and other psychometric properties of a sound assessment, it is compulsory that the items for choice be constructed comparable in item difficulty, content coverage, and especially, in latency for choice. Although this paper has systematically investigated how choices are made and how they affect performance on the basis of a paper-and-pencil test, the findings and principles established in this paper should be highly generalizable to various computer adaptive assessment modes.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Chang, R. (1988). *Chemistry.* New York: McGraw Hill.

The College Board. (1990). *The 1989 advanced placement examination in chemistry and its grading.* New York: Author.

Fremer, J., Jackson, R., & McPeek, M. (1968). *Review of the psychometric characteristics of the advanced placement tests in chemistry, American history, and French* (Unpublished Technical Memorandum). Princeton, NJ: Educational Testing Service.

Keppel, G. (1982). *Design and analysis, a researcher's handbook.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Masterton, W. L., Slowinski, E. J., & Stanitski, C. L. (1985). *Chemical principles* (6th edition). Philadelphia: Saunders College Publishing.

Pomplum, M., Morgan, R., & Nellikunnel, A. (1992). *Choice in advanced placement tests* (Unpublished Statistical Report No. 92-51). Princeton, NJ: Educational Testing Service.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? Analysis of two tests. *Journal of Educational Measurement, 31,* 113-123.

Wainer, H., & Thissen, D. (1992). *Choosing a test.* (Technical Report No. 92-25). Princeton, NJ: Educational Testing Service.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118.

Wainer, H., & Thissen, D. (1994). One test taker choice in educational testing. *Review of Educational Research, 64,* 159-195.

Wainer, H., Wang, X-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by test takers' choice? *Journal of Educational Measurement, 31,* 183-199.

Wang, X-B. (1996a, April). *Investigating allowing test takers to choose constructive response items on a test.* Paper presented at the annual meeting of the 1996 American Educational Research Association, New York.

Wang, X-B. (1996b, April). *Understanding psychological processes that underlie test takers' choices of constructed response items on a test.* Paper presented at the annual meeting of the 1996 American Educational Research Association, New York.

Wang, X-B., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow test taker choice. *Applied Measurement in Education, 8(3),* 211-225.

Wittrock, M., & Baker, E., (1992). *Testing and cognition.* Englewood Cliffs, NJ: Prentice Hall.

Appendix

Chemistry Teacher Expert Judgment Survey

Task 1: Rank the Five Essays

Introduction: 1989 AP Chemistry Examination has five essay questions in Part D, labeled as Essays 5, 6, 7, 8, and 9.

Instructions: Please rank the five essays in terms of their relative easiness and your preference to answer from (1) being the easiest essay to answer to (5) the most difficult essay to answer, and rate Questions 2-6:

Q1: Your rank: (1) _____ (2) _____ (3) _____ (4) _____ (5) _____

Q2: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 5?

(Hardly) 1 2 3 4 5 *(Extensively)*

Q3: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 6?

(Hardly) 1 2 3 4 5 *(Extensively)*

Q4: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 7?

(Hardly) 1 2 3 4 5 *(Extensively)*

Q5: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 8?

(Hardly) 1 2 3 4 5 *(Extensively)*

Q6: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 9?

(Hardly) 1 2 3 4 5 *(Extensively)*

Task 2: Qualitative Descriptions of Common Characteristics

Introduction: Students were asked to choose three of five essays. This resulted in 10 possible combinations. Based on 18,462 students who took the test in 1989, I found dramatic differences in students' choices.

Instructions: Please write about your opinions on the reasons behind such dramatic choice distributions. Attach more paper if necessary.

Question 7: Why do so many students choose essay combination 5, 6, and 8? What characteristics do they share in common?

Afterthought for Question 7 (after you work through the survey once):

Question 8: Why do so many students choose essay combination 5, 7, and 8? What characteristics do they share in common?

Afterthought for Question 8 (after you work through the survey once):

Question 9: Why do so many students choose essay combination 5, 6, and 7? What characteristics do they share in common?

Afterthought for Question 9 (after you work through the survey once):

Question 10: Why do so few students choose essay combination 6, 7, and 9? What characteristics do they share in common?

Afterthought for Question 10 (after you work through the survey once):

Question 11: Why do so moderate a number of students choose essay combination 5, 8, and 9? What characteristics do they share in common?

Afterthought for Question 11 (after you work through the survey once):

Holistic reflections or additional comments on the overall observed choice patterns.

Now you have completed your survey. If you wish, you may go over the survey again and add your afterthoughts to each of the questions you have answered.

TM034348

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").